Statistical inference and feasibility determination: a nonasymptotic approach

Zhu, Ying

University of California, San Diego

17 August 2018

Statistical inference and feasibility determination: a nonasymptotic approach*

Ying Zhu†

June 20, 2019

Abstract

We develop non-asymptotically justified methods for hypothesis testing about the $p-$dimensional coefficients θ^* in (possibly nonlinear) regression models. Given a function $h : \mathbb{R}^p \mapsto \mathbb{R}^m$, we consider the null hypothesis $H_0 : h(\theta^*) \in \Omega$ against the alternative hypothesis $H_1 : h(\theta^*) \notin \Omega$, where Ω is a nonempty closed subset of $\mathbb{R}^m$ and h can be nonlinear in θ^*. Our (nonasymptotic) control on the Type I and Type II errors holds for fixed n and does not rely on well-behaved estimation error or prediction error; in particular, when the number of restrictions in H_0 is large relative to $p - n$, we show it is possible to bypass the sparsity assumption on θ^* (for both Type I and Type II error control), regularization on the estimates of θ^*, and other inherent challenges in an inverse problem. We also demonstrate an interesting link between our framework and Farkas' lemma (in math programming) under uncertainty, which points to some potential applications of our method outside traditional hypothesis testing.

Keywords: Nonasymptotic validity; hypothesis testing; confidence regions; concentration inequalities; high dimensional regressions; Farkas' lemma

1 Introduction

A common feature of the existing procedures that are deemed "practical" for testing nonlinear hypotheses about regression coefficients is that they hinge on asymptotic validity to some extent. This occurrence is perhaps not coincidental as asymptotic

*Two earlier versions (arXiv:1808.07127v1 and arXiv:1808.07127v2) of this paper were circulated under the titles "Concentration based inference in high dimensional generalized regression models (I: statistical guarantees)" and "Concentration based inference for high dimensional (generalized) regression models: new phenomena in hypothesis testing". Substantial changes have been made in this version (arXiv:1808.07127v3). This version was prepared during my appointment at Department of Statistics and Department of Computer Science, Purdue University, West Lafayette, IN. A start-up fund from Purdue College of Science supported this research. Earlier versions were prepared during my appointment at Michigan State University (Department of Economics, Social Science Data Analytics Initiative) that also provided financial support.

†Email: yiz012@ucsd.edu. Assistant Professor of Economics (starting on July 1, 2019), Department of Economics, University of California, San Diego.

analyses often allow one to focus on some "leading" term(s) by assuming the "remainder" term(s) approach to zero faster, which can be quite convenient for determining the threshold in a test. However, many real-world applications (in controlled experiments, for example) have a limited sample size which renders any asymptotic argument questionable. Our primary goal is to find situations where effective non-asymptotic methods can be developed for hypothesis testing about the coefficients $\theta^* \in \mathbb{R}^p$ in regression models. Given a function $h : \mathbb{R}^p \mapsto \mathbb{R}^m$, let

$$H_0 : h(\theta^*) \in \Omega \text{ vs. } H_1 : h(\theta^*) \notin \Omega, \tag{1}$$

where Ω is a nonempty closed subset of $\mathbb{R}^m$ and h is allowed to be nonlinear in θ^*. Relative to existing literature, we consider these broader forms of hypotheses and the impact of m, the number of restrictions in the null hypothesis H_0. Throughout this paper, we assume that $\{\theta \in \mathbb{R}^p : h(\theta) \in \Omega\} \neq \emptyset$ and H_0 does not contain any redundant restrictions.

Our main focus is the following Gaussian regression model

$$Y_i = g(V_i; \theta^*) + W_i, \qquad i = 1, ..., n. \tag{2}$$

In the equation above, the functional form of $g(V_i; \theta^*)$ is known and may be nonlinear in θ^*; $Y = \{Y_i\}_{i=1}^n$ is an $n-$dimensional vector of responses; $V = \{V_i\}_{i=1}^n \in \mathbb{R}^{n \times k}$ is the matrix of covariates with the ith row specified by V_i; $W = \{W_i\}_{i=1}^n \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$ and is independent of V, where $\mathbf{0}_n$ denotes an $n-$dimensional vector of zeros; θ^* is a $p-$dimensional vector of unknown coefficients and p is allowed to exceed k as well as the sample size n. Throughout the paper, we make our argument by conditioning on V.

In (1), allowing h to be nonlinear functions of θ^* can be very useful. For example, policy researchers are often interested in testing whether $\frac{1}{n}\sum_{i=1}^n \frac{\partial g(v_i;\theta^*)}{\partial v_{ij}}$ or $\frac{1}{n}\sum_{i=1}^n \frac{\Delta g(v_i;\theta^*)}{\Delta v_{ij}}$ lies in some interval. These quantities are referred to as the average partial effect (APE) of v_{ij}s on $\mathbb{E}(Y_i|V_i = v_i) = g(v_i; \theta^*)$, holding all the other v_{il}s constant ($l \neq j$). For evaluating policy interventions in social science or treatment procedures in medical studies, both applied and theoretical researchers have recognized the importance of APEs for understanding the magnitudes of effects (see [21] for more discussions on this topic). If $g(v_i; \theta^*) = v_i\theta^*$ in (2), the APE of v_{ij}s is simply θ_j^*. However, when a nonlinear model is used to analyze the effects of v_{ij}s, the APEs often depend on θ^* in a nonlinear fashion and the individual coefficients themselves may no longer convey interpretable information about the effects of v_{ij}s.

This paper makes several contributions. First, the new method we propose provides a finite sample alternative to the classical asymptotic procedures (such as the Wald, Lagrange multiplier, and likelihood ratio tests) for testing a single or multiple nonlinear hypotheses about coefficients in regression models where p is small relative to n. Second, this method can be used for testing simultaneous (nonlinear) hypotheses when p is comparable to or larger than n and the number of restrictions in H_0 is large relative to $p - n$, in which case it becomes possible to

bypass the sparsity assumption on θ^*, regularization (on the estimates of θ^*), and other inherent challenges in an inverse problem. As we will see in the subsequent sections, it is quite natural that the coverage properties of our procedure as well as the control on the Type I error do not rely on any form of sparsity in θ^*. In terms of the Type II error control, we exploit that more restrictions (larger m) on θ^* in H_0 result in fewer parameters to be determined, and if m is large relative to $p-n$, it is possible for the power of our test to not rely on any form of sparsity in θ^*. This result suggests that our procedure can be useful for many economic applications where the ratio $\frac{p}{n}$ often stays constant but below 1, θ^* may not have any sparsity structure, and the null hypothesis contains nonlinear restrictions (like the APE example given previously).

Third, we provide a different interpretation of our approach by exploring a pair of primal and dual optimization problems along with the so-called Farkas's lemma (see, e.g., [2]). In particular, the primal (dual) problem is a feasibility (minimization) program where some or all of entries in the target (cost) vector are "corrupted" with additive i.i.d. noise, e.g., Gaussian. This perspective provides an interesting connection between our framework and Farkas' lemma (in math programming) under uncertainty, which points to some potential applications of our method outside traditional hypothesis testing.

1.1 Hypothesis testing in regression models

We choose our test statistics in the form of

$$\Psi_q(\hat{\theta}_\alpha) := \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - g\left(V_i; \hat{\theta}_\alpha\right) \right] \right\|_q \tag{3}$$

for some pre-specified $L-$dimensional vector of functions $f(V_i) =: X_i$. In (3), $\hat{\theta}_\alpha$ is obtained by solving the following program[1]:

$$\left(\hat{\theta}_\alpha,\, \hat{\mu}_\alpha\right) \in \arg \min_{(\theta_\alpha, \mu_\alpha) \in \mathbb{R}^p \times \mathbb{R}} \mu_\alpha$$

$$\text{subject to: } \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - g\left(V_i; \theta_\alpha\right) \right] \right\|_q \leq r_{\alpha,q} + \mu_\alpha, \tag{4}$$

$$h(\theta_\alpha) \in \Omega, \tag{5}$$

$$\mu_\alpha \geq 0, \tag{6}$$

with $q \in [1, \infty]$ chosen by the users. For $1 \leq q \leq \infty$, we write $\|\Delta\|_q$ to mean the l_q-norm of an $L-$dimensional vector Δ, where $\|\Delta\|_q := \left(\sum_{i=1}^{L} |\Delta_i|^q\right)^{1/q}$ when $1 \leq q < \infty$ and $\|\Delta\|_q := \max_{i=1,\ldots,L} |\Delta_i|$ when $q = \infty$. Here we suppress the dependence of $\left(\hat{\theta}_\alpha,\, \hat{\mu}_\alpha\right)$ in (4) on q for notational simplicity. The choice for $r_{\alpha,q}$ in

[1] Section A.1 provides an alternative formulation.

the first set of constraints is to be specified in Section 2. Statistical guarantees in this paper are stated in terms of (α, q).

Suppose $g(V_i; \theta^*) = V_i\theta^*$ in (2). For the special case $q = \infty$ and $X_i = V_i$, it is worth comparing program (4) with the well-known Dantzig selector:

$$\hat{\theta}^{dan} \in \arg\min_{\theta \in \mathbb{R}^p} \|\theta\|_1 \quad \text{subject to} \left\| \frac{1}{n} V^T (Y - V\theta) \right\|_\infty \leq r. \tag{7}$$

Unlike (7), (4) involves a slack variable μ_α in the first set of constraints, (5) and (6), as well as a different objective function (minimizing the slack variable, instead of minimizing $\|\theta\|_1$). Consequently, the resulting solution to (7) is not constrained to satisfy $h(\hat{\theta}^{dan}) \in \Omega$, whereas $\hat{\theta}_\alpha$ in (4) satisfies $h(\hat{\theta}_\alpha) \in \Omega$.

A solution $\hat{\theta}_\alpha$ to (4) may not necessarily be unique: that is, there might be different $\hat{\theta}_\alpha$s that satisfy (4) while delivering the same (minimal) objective value $\hat{\mu}_\alpha$. We refer to the non-negative scalar μ_α in (4) as the "slack" variable that fills the gap between $\left\| \frac{1}{n} \sum_{i=1}^n X_i [Y_i - g(V_i; \theta^*)] \right\|_q$ and $\left\| \frac{1}{n} \sum_{i=1}^n X_i [Y_i - g(V_i; \theta_\alpha)] \right\|_q$ where $h(\theta_\alpha) \in \Omega$. When the null hypothesis is true, i.e., $h(\theta^*) \in \Omega$, the optimal value $\hat{\mu}_\alpha$ must be zero with probability at least $1 - \alpha$.

We derive (non-asymptotic) thresholds $r_{\alpha,q}$ such that

$$\mathbb{P}_0 \left\{ \Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q} \right\} \leq \alpha, \qquad \text{(Type I Error)} \tag{8}$$

$$\mathbb{P}_1 \left\{ \Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q} \right\} \leq \beta, \qquad \text{(Type II Error)} \tag{9}$$

where $\mathbb{P}_0$ means under H_0, $\mathbb{P}_1$ means under H_1 and a "Level β Separation Requirement" (to be introduced in Section 2.1.3). Our decision rule is that if $\Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q}$, we reject the null hypothesis H_0 in (1) at the $1 - \alpha$ confidence level. In addition to the guarantees on the Type I and Type II errors, we also construct confidence regions in terms of $\left(\hat{\theta}_\alpha, \hat{\mu}_\alpha\right)$.

1.2 Comparison with existing results

At first glance, there seems no lack of non-asymptotic bounds on the l_p-error (often $p \in [1, 2]$ or $p = \infty$) of some regularized estimator (such as the various Lasso estimators and Dantzig selectors) concerning a sparse linear regression model (e.g., [5]). However, these bounds (even in the sharpest forms) tend to involve quite a few unknown nuisance parameters that are hard to estimate in practice. In order to adapt the existing bounds for the purpose of inference, knowledge on the sparsity of θ^* would be needed at a minimum; see the discussion in [10]. By contrast, the nonasymptotic thresholds in our testing procedure do not involve any unknown parameters related to sparsity.

In particular, this paper studies nonasymptotic inference by exploiting the concentration phenomenon, which should be distinguished from another line of literature based on normal approximations using the Stein's Method, for example, [8],

[11] (also see [24, 25] whose methods are justified by the theory in [8]). In particular, [11] develops results for hypothesis testing in the regime of $n \gg p$; by contrast, we also study the regime where p is comparable to or larger than n. In [11], some of the results are still only asymptotically valid and the other results (even though nonasymptotically justified) come with probabilistic guarantees that contain rather loose constants and dimension-dependent components.

If $g\left(V_i; \theta^*\right) = V_i\theta^*$ in (2) and we choose $q = \infty$, $X_i = V_i$, then (3) is reduced to

$$\Psi_\infty\left(\hat{\theta}_\alpha\right) := \left\| \frac{1}{n} \sum_{i=1}^{n} V_i \left(Y_i - V_i\hat{\theta}_\alpha\right) \right\|_\infty .$$

This statistics shares some resemblance to the score-based correction term in the debiased Lasso literature (see, e.g., [9, 12, 19, 23, 24]) as well as the decorrelated score in [15]. Unlike the debiased and decorrelated procedures which require an initial (consistent) estimator for (the sparse) θ^* in the correction term, our $\hat{\theta}_\alpha$ here need not be consistent and is directly used in the test statistics (requiring no further debiasing or decorrelating step). In addition, our method is nonasymptotically valid and does not require θ^* to have any sparse structure, whereas the aforementioned papers hinge on the asymptotic normality of the debiased or decorrelated procedure and require θ^* to be sufficiently sparse.

For choosing the non-asymptotic thresholds $r_{\alpha,q}$ for general $q \in [1, \infty]$, our proposal exploits sharp concentration of Lipschitz functions of Gaussian variables. The key component in our $r_{\alpha,q}$ is data-driven and uses a Monte-Carlo approximation to "mimic" the expectation that is concentrated around and automatically captures the dependencies across coordinates. In this perspective, our results share some similarity as those in [1]; however, [1] concern inference for the mean of a random vector while we consider inference about the coefficients ($\theta^* \in \mathbb{R}^p$) in regression models.

In terms of relaxing sparsity assumptions, this paper shares slight similarity as [25] and [26] although our focus and approach are drastically different from those in [25] and [26]. First, [25] deal with $H_0 : \theta_j^* = \theta_j^0$ for $j \in M \subseteq \{1, ..., p\}$ and [26] deal with $H_0 : a^T\theta^* = b^0$ for some prespecified $a \in \mathbb{R}^p$, $b^0 \in \mathbb{R}$. Second, parts of the analyses in [25] and [26] still rely on sparsity to some extent (albeit considerably less compared to much of the existing literature) even in cases where $\frac{p}{n}$ stays constant but below 1; as a consequence, their methods still involve regularization (on the estimates of subcomponents of θ^*). Finally, like much of the existing literature, the statistical guarantees in [25] and [26] are asymptotic and require $n \to \infty$.

1.3 Feasibility determination under uncertainty

The proposed framework in this paper can be motivated with a classical notion called "certificate of infeasibility" in operations research. Essentially, we can view hypothesis testing as determining a given system of inequalities is feasible or not when some of the system inputs are subject to uncertainty. For the exposition in

this section, let us introduce the following pair of problems:

$$\begin{aligned} \max_{\theta \in \mathbb{R}^p} \quad & \mathbf{0}_p^T \theta \\ \text{subject to} \quad & A\theta = b \\ & \theta \geq 0 \end{aligned} \tag{10}$$

and

$$\begin{aligned} \min_{\pi \in \mathbb{R}^d} \quad & \pi^T b \\ \text{subject to} \quad & \pi^T A \geq \mathbf{0}_p^T. \end{aligned} \tag{11}$$

The book [2] refers to the $d-$dimensional column vectors in the matrix $A \in \mathbb{R}^{d \times p}$ as the resource vectors, $b \in \mathbb{R}^d$ as the target vector in (10) and the cost vector in (11); $\theta \in \mathbb{R}^p$ and $\pi \in \mathbb{R}^d$ are called the vectors of decision variables to (10) and (11), respectively. Note that program (10) is the dual of program (11) and vice versa. By convention in linear programming, the constraints in (10) are called the "standard form" and the d rows of A are assumed to be linearly independent (so we must have $d \leq p$). Algorithms designed for linear programming can be used to determine whether a feasible solution exists for (10).

The pair (10)-(11) is associated with the famous Farkas' lemma, which states that: (1) if (10) is infeasible, then there exists some π feasible for (11) such that $\pi^T b < 0$; (2) if (10) is feasible, then there cannot exist a π feasible for (11) such that $\pi^T b < 0$. This lemma has an important application in asset pricing, in particular the so called "absence of arbitrage" condition that underlies much of finance theory (see the example in [2]).

In the classical paradigm, b is assumed to be precisely known and not subject to uncertainty. Started in the early 1970s, developing robust approaches to solving linear programming problems where data are subject to uncertainty has been extensively studied. Some of the well-known papers include, for example, [3, 4, 18]. Instead of contributing new optimization theory to this literature, our goal here is to show an interesting link between our framework and Farkas' lemma under uncertainty. Let us consider a special case of uncertainty where the first $n\,(\leq d)$ components in b are corrupted with i.i.d. Gaussian noise, in which scenario we only observe $Y_i = b_i + W_i$ with $W_i \sim \mathcal{N}(0,\, \sigma^2)$ for $i = 1, ..., n$. We are interested in the following pairs of hypotheses:

$$\begin{aligned} H_0 &: \text{A solution to (10) exists} \\ H_1 &: \text{No solution exists for (10),} \end{aligned} \tag{12}$$

which is equivalent to

$$\begin{aligned} H_0 &: \text{There cannot exist a } \pi \text{ feasible for (11) such that } \pi^T b < 0 \\ H_1 &: \text{There exists a } \pi \text{ feasible for (11) such that } \pi^T b < 0. \end{aligned} \tag{13}$$

To test the hypotheses above, we can adapt (4) and solve the following program:

$$\begin{aligned}
\left(\hat{\theta}_\alpha,\, \hat{\mu}_\alpha\right) \in \arg \min_{(\theta_\alpha, \mu_\alpha) \in \mathbb{R}^p \times \mathbb{R}} & \mu_\alpha \\
\text{subject to: } \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left(Y_i - V_i \theta_\alpha\right) \right\|_q & \leq r_{\alpha,q} + \mu_\alpha, \\
A_i \theta_\alpha &= b_i, i = n+1, ..., d, \\
\theta_\alpha &\geq 0, \\
\mu_\alpha &\geq 0,
\end{aligned} \tag{14}$$

for some pre-specified $L-$dimensional vector of functions $f\left(V_i\right) =: X_i$, where $V_i = A_i$ for $i = 1, ..., n$ and A_i is the ith row of A. Note that program (14) is simply a special instance of (4), where the constraints $h(\theta_\alpha) \in \Omega$ correspond to

$$\left\{\theta_\alpha \geq 0,\, A_i \theta_\alpha = b_i,\, i = n+1, ..., d\right\}.$$

Letting

$$\Psi_q(\hat{\theta}_\alpha) := \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left(Y_i - V_i \hat{\theta}_\alpha\right) \right\|_q, \tag{15}$$

if $\Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q}$, we then reject (12) and (13) at the $1-\alpha$ confidence level. As in Section 1.1, (8) and (9) also hold for (15). This result can be viewed as the Farkas' lemma extended to cases where the target vector b in (10) (respectively, the cost vector in (11)) is corrupted with i.i.d. noise.

1.4 Organization of this paper

Section 2 establishes nonasymptotic control on the Type I and Type II errors for our method. We provide numerical evidence through a simulation study in Section 3. Section 4 provides additional extensions, which include some nonasymptotic justifications for inference in statistical models that involve non-Gaussian responses. All technical details are collected in the supplementary materials.

2 Main results

For the regression model (2), we begin with the scenario where σ^2 is known, and then consider the scenario where σ^2 is unknown. Throughout this section, we use $\mathbb{E}_W[\cdot]$ to denote the expectation over W only, conditioning on V.

Recalling the pre-specified $L-$dimensional vector of functions $f\left(V_i\right) =: X_i$ in Section 1.1, our first result establishes an "ideal" confidence region for

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \hat{\theta}^*_\alpha\right) \right] \right\|_q.$$

The "theoretical" optimal solution above, $\hat{\theta}^*_\alpha$, is obtained by setting $r_{\alpha,q}$ in (4) to $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]$ plus a deviation. In practice, $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]$ may be bounded with its Monte Carlo approximation and a "small" deviation term. This approach results in a "practical" optimal solution, $\hat{\theta}_\alpha$, which can then be used to construct test statistics and a "practical" confidence region.

To state the first result, we introduce the following notation (which will appear in many places throughout this paper):

$$\begin{aligned}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q &= \sqrt[q]{\sum_{j=1}^L\left(\sqrt{\frac{1}{n}\sum_{i=1}^n X_{ij}^2}\right)^q}, \qquad q\in[1,\,\infty)\\ \left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q &= \max_{j\in\{1,\ldots,L\}}\sqrt{\frac{1}{n}\sum_{i=1}^n X_{ij}^2}, \qquad q=\infty.\end{aligned}$$

We can assume X is normalized such that $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = 1$ for all $j = 1, ..., L$, in which case $\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q = L^{\frac{1}{q}}$ for $q \in [1,\,\infty)$ and $\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q = 1$ for $q = \infty$.

Proposition 2.1. *Assume (2) where $W \sim \mathcal{N}(\mathbf{0}_n,\,\sigma^2\boldsymbol{I}_n)$ and is independent of V. Then for any $q \in [1,\,\infty]$, we have*

$$\mathbb{P}\left\{\left\|\frac{1}{n}X^TW\right\|_q \geq \mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right] + t\right\} \leq \exp\left(\frac{-nt^2}{2\sigma^2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q^2}\right). \tag{16}$$

Moreover, for $\alpha \in (0,\,1)$, let

$$r_{\alpha,q} = r^*_{\alpha,q} := \mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right] + \sigma\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q\sqrt{\frac{2}{n}\log\frac{1}{\alpha}} \tag{17}$$

in (4). Then, an optimal solution $\left(\hat{\theta}^*_\alpha,\,\hat{\mu}^*_\alpha\right)$ to (4) must satisfy

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[g\left(V_i;\theta^*\right) - g\left(V_i;\hat{\theta}^*_\alpha\right)\right]\right\|_q \geq \hat{\mu}^*_\alpha, \tag{18}$$

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[g\left(V_i;\theta^*\right) - g\left(V_i;\hat{\theta}^*_\alpha\right)\right]\right\|_q \leq 2r^*_{\alpha,q} + \hat{\mu}^*_\alpha, \tag{19}$$

with probability at least $1 - \alpha$.

2.1 Hypothesis testing

For the moment, suppose we set $r_{\alpha,q} = r^*_{\alpha,q}$ in (4) according to (17). Under H_0, $(\theta^*, 0)$ is an optimal solution to (4). Consequently, given the test statistics (3) and a chosen $\alpha \in (0, 1)$, any optimal solution to (4) must satisfy

$$\mathbb{P}_0 \left\{ \Psi_q(\hat{\theta}^*_\alpha) \geq r^*_{\alpha,q} \right\} \leq \alpha \tag{20}$$

where $\mathbb{P}_0$ means under H_0.

The claim in (20) suggests a test (with level α) based on the statistics $\Psi_q(\hat{\theta}^*_\alpha)$ and an "ideal" critical value, $r^*_{\alpha,q}$, given in (17). When $W \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$ and σ^2 is known, the first term $\mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right]$ in $r^*_{\alpha,q}$ can be approximated by Monte-Carlo as follows. Let $Z \in \mathbb{R}^{n \times R}$ (independent of W and V) be a matrix consisting of independent entries randomly drawn from $\mathcal{N}(0, 1)$ and the rth column of Z is denoted by Z_r. By (63) and (64) in Section A.2, note that $\sigma R^{-1} \sum_{r=1}^R \left\| \frac{1}{n} X^T Z_r \right\|_q - \mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right]$ is sub-Gaussian with parameter at most $(nR)^{-1/2} \sigma \left\| \sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2} \right\|_q$. Consequently, (61) yields the following concentration

$$\mathbb{P} \left\{ \mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] \geq \frac{\sigma}{R} \sum_{r=1}^R \left\| \frac{1}{n} X^T Z_r \right\|_q + t \right\} \leq \exp \left(\frac{-nRt^2}{2\sigma^2 \left\| \sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2} \right\|_q^2} \right). \tag{21}$$

Combining (16) and (21) yields

$$\begin{aligned} &\mathbb{P} \left\{ \left\| \frac{1}{n} X^T W \right\|_q \geq \frac{\sigma}{R} \sum_{r=1}^R \left\| \frac{1}{n} X^T Z_r \right\|_q + t_1 + t_2 \right\} \\ \leq &\exp \left(\frac{-nt_1^2}{2\sigma^2 \left\| \sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2} \right\|_q^2} \right) + \exp \left(\frac{-nRt_2^2}{2\sigma^2 \left\| \sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2} \right\|_q^2} \right). \end{aligned} \tag{22}$$

2.1.1 Construction of critical values ($r_{\alpha,q}$) and Type I error

For some chosen α_1, $\alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = \alpha \in (0, 1)$, we let in (22),

$$\begin{aligned} t_1 &= \sigma \left\| \sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2} \right\|_q \sqrt{\frac{2}{n} \log \frac{1}{\alpha_1}} := \tau_{\alpha_1,q}, \\ t_2 &= \sigma \left\| \sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2} \right\|_q \sqrt{\frac{2}{nR} \log \frac{1}{\alpha_2}} := \sqrt{\frac{1}{R}} \tau_{\alpha_2,q}. \end{aligned} \tag{23}$$

Based on (22) along with the choices of t_1 and t_2 above, we set in (4),

$$r_{\alpha,q} = \frac{\sigma}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}X^T Z_r\right\|_q + \tau_{\alpha_1,q} + \sqrt{\frac{1}{R}}\tau_{\alpha_2,q}. \tag{24}$$

Note that we can draw as many columns in Z as we want, to make $\sqrt{\frac{1}{R}}\tau_{\alpha_2,q}$ in (24) small; for a given α, we can let α_2 be smaller than α_1 because of the additional "$\sqrt{\frac{1}{R}}$".

Under H_0, $(\theta^*, 0)$ is an optimal solution to (4) with $r_{\alpha,q}$ specified in (24). Consequently, a (practical) optimal solution to (4) must satisfy

$$\mathbb{P}_0\left\{\Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q}\right\} \leq \alpha \qquad \textit{(Type I Error).} \tag{25}$$

Remarks. In terms of control on the Type I error, the l_q-norm in (3) and (4) can be generalized to the function $\zeta_q : \mathbb{R}^L \mapsto \mathbb{R}$ that satisfies:

- for all $z \in \mathbb{R}^L$ and $a \in \mathbb{R}^+$, $\zeta_q(az) = a\zeta_q(z)$,
- for all $z,\, z^{'} \in \mathbb{R}^L$, $\zeta_q\left(z + z^{'}\right) \leq \zeta_q(z) + \zeta_q(z^{'})$,
- for all $z \in \mathbb{R}^L$, $|\zeta_q(z)| \leq \|z\|_q$ for $q \in [1, \infty]$.

This generalization is inspired by the construction in [1]. We simply let

$$r_{\alpha,q} = \frac{\sigma}{R}\sum_{r=1}^{R}\zeta_q\left(\frac{1}{n}X^T Z_r\right) + \tau_{\alpha_1,q} + \sqrt{\frac{1}{R}}\tau_{\alpha_2,q},$$

and obtain

$$\mathbb{P}_0\left\{\zeta_q\left(\frac{1}{n}\sum_{i=1}^{n}X_i\left[Y_i - g\left(V_i;\hat{\theta}_\alpha\right)\right]\right) \geq r_{\alpha,q}\right\} \leq \alpha \qquad \text{(Type I Error)},$$

where $\hat{\theta}_\alpha$ is a solution to (4) with the l_q-norm in the first set of constraints replaced by ζ_q. Given ζ_q is subadditive and bounded by the l_q-norm, the result above follows from the simple fact that

$$\begin{aligned}\left|\zeta_q\left(\frac{1}{n}X^T W\right) - \zeta_q\left(\frac{1}{n}X^T W^{'}\right)\right| &\leq \left\|\frac{1}{n}X^T\left(W - W^{'}\right)\right\|_q \\ &\leq \frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q \left\|W - W^{'}\right\|_2.\end{aligned}$$

Consequently, we can establish bounds that are identical to (16), (21), (22) in terms of $\zeta_q\left(\frac{1}{n}X^T W\right)$, $\mathbb{E}_W\left[\zeta_q\left(\frac{1}{n}X^T W\right)\right]$, $\frac{\sigma}{R}\sum_{r=1}^{R}\zeta_q\left(\frac{1}{n}X^T Z_r\right)$, and then follow the same argument as what is used to show (25).

2.1.2 Practical confidence regions

Let $\left(\hat{\theta}_\alpha,\, \hat{\mu}_\alpha\right)$ be an optimal solution to (4) with $r_{\alpha,q}$ specified in (24). We have

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \hat{\theta}_\alpha\right) \right] \right\|_q \\ \leq & \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - g\left(V_i; \hat{\theta}_\alpha\right) \right] \right\|_q + \left\| \frac{1}{n} X^T W \right\|_q \\ \leq & \frac{2\sigma}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} X^T Z_r \right\|_q + 2\tau_{\alpha_1,q} + 2\sqrt{\frac{1}{R}} \tau_{\alpha_2,q} + \hat{\mu}_\alpha \end{aligned} \tag{26}$$

and

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \hat{\theta}_\alpha\right) \right] \right\|_q \geq \hat{\mu}_\alpha \tag{27}$$

with probability at least $1 - \alpha$. The argument for (27) is identical to what is used to show (18) (see more details in Section A.3). As we have pointed out in the introduction, there might be different $\hat{\theta}_\alpha$s that solve (4) while producing the same (minimal) objective value $\hat{\mu}_\alpha$. Consequently, there may be more than one confidence interval in the form of (26) and (27). In view of these intervals, the length of them is naturally

$$CI - Length = \frac{2\sigma}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} X^T Z_r \right\|_q + 2\tau_{\alpha_1,q} + 2\sqrt{\frac{1}{R}} \tau_{\alpha_2,q}. \tag{28}$$

If $\mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right]$ can be known exactly and we were able to set $r_{\alpha_1,q} = r^*_{\alpha_1,q}$ in (4) as in Proposition 2.1, then any resulting optimal solution $\left(\hat{\theta}^*_\alpha,\, \hat{\mu}^*_\alpha\right)$ to (4) should satisfy

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \hat{\theta}^*_\alpha\right) \right] \right\|_q \leq 2\mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] + 2\tau_{\alpha_1,q} + \hat{\mu}^*_\alpha \tag{29}$$

with probability at least $1 - \alpha_1$. Comparing (26) with (29), note that the difference in the right hand sides is

$$2 \left(\frac{\sigma}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} X^T Z_r \right\|_q - \mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] \right) + 2\sqrt{\frac{1}{R}} \tau_{\alpha_2,q},$$

which can be made arbitrarily small with a large number of random draws in the Monte-Carlo approximation. Because of such an approximation, the probabilistic guarantee for (26) is bounded from below by $1 - \alpha$ instead of $1 - \alpha_1$.

Given the statistics $\Psi_q(\hat{\theta}_\alpha)$ in (3) based on (a practical) $\hat{\theta}_\alpha$ and the critical value $r_{\alpha,q}$ defined in (24), we have constructed a test with level α as shown in (25). For a

chosen $\beta \in (0,\, 1)$, when can this test correctly detect an alternative with probability at least $1-\beta$? To answer this question, we introduce the "Separation Requirement" in the following section.

2.1.3 Separation requirement and Type II error

Letting $\Theta_0 := \{\theta \in \mathbb{R}^p : h(\theta) \in \Omega\}$, we choose $\beta_1,\, \beta_2 > 0$ such that $\beta_1 + \beta_2 = \beta \in (0,\, 1)$, and assume

$$\inf_{\theta \in \Theta_0} \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \theta\right) \right] \right\|_q \geq \delta_{\alpha,\beta,q} \tag{30}$$

with

$$\delta_{\alpha,\beta,q} = 2\mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] + \tau_{\alpha_1,q} + \sqrt{\frac{1}{R}} \tau_{\alpha_2,q} + \sqrt{\frac{1}{R}} \tau_{\beta_1,q} + \tau_{\beta_2,q} \tag{31}$$

for the prespecified $\alpha_1,\, \alpha_2 > 0$ (used in (24)) such that $\alpha_1 + \alpha_2 = \alpha \in (0,\, 1)$. We will refer to (30) as the "Separation Requirement" at the level β.

Our next result concerns the Type II error of the test based on $\Psi_q(\hat{\theta}_\alpha)$ in (3) and $r_{\alpha,q}$ defined in (24). For completeness, we also include the result for the Type I error.

Theorem 2.1. *Assume (2) where $W \sim \mathcal{N}(\mathbf{0}_n,\, \sigma^2 \boldsymbol{I}_n)$ and is independent of V. For chosen $\alpha_1,\, \alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = \alpha \in (0,\, 1)$, consider the statistics $\Psi_q(\hat{\theta}_\alpha)$ based on $\hat{\theta}_\alpha$ and the critical value $r_{\alpha,q}$ defined in (24). For any $q \in [1,\, \infty]$, we have*

$$\mathbb{P}_0 \left\{ \Psi_q(\hat{\theta}_\alpha) \geq r_{\alpha,q} \right\} \leq \alpha, \qquad \textit{(Type I Error)} \tag{32}$$

where $\mathbb{P}_0$ means under H_0. For the same $r_{\alpha,q}$ used in (32) and chosen $\beta_1,\, \beta_2 > 0$ such that $\beta_1 + \beta_2 = \beta \in (0,\, 1)$, if $h(\theta^*) \notin \Omega$ and (30) is satisfied, we have

$$\mathbb{P}_1 \left\{ \Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q} \right\} \leq \beta, \qquad \textit{(Type II Error)} \tag{33}$$

where $\mathbb{P}_1$ means under H_1 and (30).

2.1.4 Discussions of the results

Some observations can be made from the results we have established so far. First, our guarantees do not rely on any form of sparsity in θ^*, well behaved $\left\| \hat{\theta}_\alpha - \theta^* \right\|_2$ or $\sqrt{\frac{\sum_{i=1}^{n} \left[g(V_i;\theta^*) - g\left(V_i;\hat{\theta}\right) \right]^2}{n}}$.

Second, the number of restrictions (i.e., m) in H_0 plays a significant role in our separation requirement. Suppose $\Theta_0 \neq \emptyset$. For (30) to hold, it is almost necessary that $m > p - n$. If $p \leq n$, this "necessary" condition is satisfied for any $m > 0$.

If $p > n$ but $m \leq p - n$, we can always find a solution $\hat{\theta}$ such that $h(\hat{\theta}) \in \Omega$ and $g\left(V_i; \hat{\theta}\right) = Y_i$ for all i. Consequently, we have

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \hat{\theta}\right) \right] \right\|_q \\
= &\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - g\left(V_i; \hat{\theta}\right) \right] - \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - g\left(V_i; \theta^*\right) \right] \right\|_q = \left\| \frac{1}{n} X^T W \right\|_q . \qquad (34)
\end{aligned}$$

By (16),

$$\mathbb{P}\left(\left\| \frac{1}{n} X^T W \right\|_q \leq \mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] + \tau_{\alpha,q} \right) \geq 1 - \alpha.$$

Since $\delta_{\alpha,\beta,q} > \mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right] + \tau_{\alpha,q}$, for a reasonably small Type I error $\alpha > 0$, $\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \hat{\theta}\right) \right] \right\|_q \geq \delta_{\alpha,\beta,q} \right)$ is small and our procedure is unlikely to detect the alternatives. As m gets larger relative to $p - n$, it becomes easier for (30) to be satisfied and our procedure to detect the alternative. Note that $\left(\hat{\theta},\, 0\right)$ also solves (4) with probability 1 for any $r_{\alpha,q} \geq 0$, and clearly the control on the Type I error remains valid.

It is instructive to consider the case $q = \infty$. Whenever $p - n < m < p$, for (30) to hold, it is almost necessary that $L + m > p$ (recalling L is the dimension of X_i). Suppose $L + m \leq p$. We can always find a solution $\hat{\theta}$ such that $\frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - g\left(V_i; \hat{\theta}\right) \right] = \mathbf{0}_L$ and $h(\hat{\theta}) \in \Omega$. Once again, this fact leads to (34) and a similar consequence.

Letting $\hat{\theta}_\alpha$ be a solution that gives the optimal value $\hat{\mu}_\alpha$ and the $L-$dimensional vector

$$\begin{aligned}
\tilde{\mu}_\alpha &= \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - g\left(V_i; \hat{\theta}_\alpha\right) \right] - \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - g\left(V_i; \theta^*\right) \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \hat{\theta}_\alpha\right) \right],
\end{aligned}$$

Section A.3 then implies that $\left(\hat{\theta}_\alpha,\, \left\|\tilde{\mu}_\alpha\right\|_\infty\right)$ is a feasible solution to (4) with probability at least $1 - \alpha$. Consequently, the optimality of $\left(\hat{\theta}_\alpha,\, \hat{\mu}_\alpha\right)$ implies

$$\left\|\tilde{\mu}_\alpha\right\|_\infty = \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \hat{\theta}_\alpha\right) \right] \right\|_\infty \geq \hat{\mu}_\alpha.$$

In summary, as $\hat{\mu}_\alpha$ increases, the actual separation $\left\|\tilde{\mu}_\alpha\right\|_\infty$ will never decrease and it becomes easier for our procedure to detect the alternatives.

This fact has an implication on the choices of $X_i = f\left(V_i\right)$, which can be quite flexible in (4) given W_i is independent of V_i. Note that including more components

in X_i may or may not increase the optimal value $\hat{\mu}_\alpha$ as larger L introduces more constraints in (4) but at the same time increases $r_{\alpha,\infty}$. However, the latter effect might be dominated by the former when the increase in L is moderate: as we will see in (40), one can show that $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] \precsim \sqrt{\frac{\log L}{n}}$, and by (67) of Section A.4,

$$r_{\alpha,\infty} \precsim \sqrt{\frac{\log L}{n}} + \sigma\sqrt{\frac{2}{n}\log\frac{1}{\alpha_1}} + \sigma\sqrt{\frac{2}{nR}\log\frac{1}{\alpha_2}} \tag{35}$$

with high probability, where we have imposed the normalization on X such that $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = 1$ for all $j = 1, ..., L$. Because of the "$\log L$" factor in (35), under the alternative hypothesis, a somewhat larger L may shrink the feasible region for (4) and increase the optimal value $\hat{\mu}_\alpha$ as well as the actual separation $\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[g\left(V_i;\theta^*\right) - g\left(V_i;\hat{\theta}_\alpha\right)\right]\right\|_\infty$.

2.2 The union bound alternative

As an alternative, the union bound approach can also be used to construct a testing procedure. In particular, we can solve (4) with $q = \infty$ and

$$r_{\alpha,\infty} = \sqrt{\max_{j\in\{1,...,L\}}\frac{2\sigma^2}{n}\sum_{i=1}^n X_{ij}^2}\sqrt{\frac{1}{n}\log\frac{2L}{\alpha}}. \tag{36}$$

Consequently, the separation distance in (30) that allows us to correctly detect an alternative with probability at least $1-\beta$ takes the form

$$\begin{aligned}\delta_{\alpha,\beta,\infty} &= r_{\alpha,\infty} + r_{\beta,\infty} \\ &= \sqrt{\max_{j\in\{1,...,L\}}\frac{2\sigma^2}{n}\sum_{i=1}^n X_{ij}^2}\left(\sqrt{\frac{1}{n}\log\frac{2L}{\alpha}} + \sqrt{\frac{1}{n}\log\frac{2L}{\beta}}\right).\end{aligned} \tag{37}$$

Given (36) and (37), we can apply the same argument for Theorem 2.1 and arrive at (32) and (33) (with $q = \infty$ in both). In contrast to our previous concentration approach, (36) is derived from a simple union bound on $\left\|\frac{1}{n}X^TW\right\|_\infty$; as a consequence, the resulting threshold $r_{\alpha,\infty}$ does not capture the dependencies between the coordinates.

We observe from (23), (24) and (31) that the quantities taking the form of $\sqrt{\log\frac{1}{\varsigma}}$ are dimension free in the concentration approach. Instead, the leading term $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right]$ in (16) and (31) reflects the "dimension complexity" and automatically takes into consideration the dependencies between the coordinates. This result is a direct consequence of the concentration phenomenon in Lipschitz functions of Gaussians. Suppose X is normalized such that $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = 1$ for all $j = 1, ..., L$, and $\tau_{\varsigma,\infty} = \sigma\sqrt{\frac{2}{n}\log\frac{1}{\varsigma}}$. Take $q = \infty$, $W \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{I}_n)$ and consider the extreme

example where X consists of L copies of the same column X_0. Then, we have

$$\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] = \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}}$$

and (31) becomes

$$\delta_{\alpha,\beta,\infty} = 2\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}} + \tau_{\alpha_1,\infty} + \sqrt{\frac{1}{R}}\tau_{\alpha_2,\infty} + \sqrt{\frac{1}{R}}\tau_{\beta_1,\infty} + \tau_{\beta_2,\infty}, \tag{38}$$

which involves no dimension complexity (as desired).

Beyond the extreme example, more generally for $q = \infty$ and $W \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{I}_n)$ (without much loss of generality by assuming $\sigma = 1$), Section A.5 shows that

$$\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] \geq \frac{1}{2}\left(1 - \frac{1}{e}\right)\sqrt{\frac{\log L}{4n^2}\min_{j,l\in\{1,\ldots,L\}}\sum_{i=1}^n (X_{ij} - X_{il})^2} \tag{39}$$

for all $L \geq 20$, and

$$\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] \leq \sqrt{\frac{2\log L}{n^2}\max_{j\in\{1,\ldots,L\}}\sum_{i=1}^n X_{ij}^2} + \sqrt{\frac{8}{n^2\log L}\max_{j\in\{1,\ldots,L\}}\sum_{i=1}^n X_{ij}^2} \tag{40}$$

for all $L \geq 2$. While the nonasymptotic validity of our testing procedure does not require any growth conditions on the dimensionality, we see from (39) that $\delta_{\alpha,\beta,\infty}$ can tend to zero only when $\frac{\log L}{n} = o(1)$ (if X does not contain identical columns).

In certain situations, (24) and (31) (with $q = \infty$ in both) can be more conservative than their union bound counterparts. On the other hand, the extreme example discussed previously suggests that in situations where X consists of realizations from highly dependent random vectors, $r_{\alpha,\infty}$ and $\delta_{\alpha,\beta,\infty}$ from the union bound approach can be bigger than (24) and (31) (with $q = \infty$ in both), respectively, due to the extra "$\log L$" term. This result is similar in spirit to those of [1], which study bootstrap confidence regions for the mean of a random vector with the concentration approach. For implementation of (4) with $q = \infty$, we may set $r_{\alpha,\infty}$ to the minimum of (24) (with $q = \infty$) and (36).

2.3 Unknown noise variance

When no prior information on σ is available, $\sqrt{\mathrm{Var}(Y_i)}$ may be used as an upper bound. We can easily estimate $\sqrt{\mathrm{Var}(Y_i)}$ by $\hat{\sigma}_Y = \sqrt{n^{-1}\sum\left(Y_i - \bar{Y}\right)^2}$ where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. Proposition 4.1 in [1] implies that

$$\sqrt{\mathrm{Var}(Y_i)} \leq \left(C_n - \frac{1}{\sqrt{n}}\Phi^{-1}\left(\frac{\kappa}{2}\right)\right)^{-1}\hat{\sigma}_Y := \bar{B}_\kappa$$

with probability at least $1 - \kappa$, where $C_n = \sqrt{\frac{2}{n}}\frac{\Gamma(n/2)}{\Gamma((n-1)/2)} = 1 - O(n^{-1})$.

In problems where $\text{Var}(W_i)$ is a constant over i, V is fixed, and the only source of randomness in Y comes from W, replacing σ with $\bar{B}_\kappa$ does not make $r_{\alpha,q}$ a more conservative threshold for constructing confidence regions. In problems with a random design, using $\bar{B}_\kappa$ could result in confidence regions that are more conservative. We note that it is rather challenging to estimate σ precisely and obtain a sharp threshold simultaneously within our framework. The main reason is that our guarantee does not require a small $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[g\left(V_i;\theta^*\right)-g\left(V_i;\hat{\theta}\right)\right]^2}$ with high probability, which seems to be needed for consistent estimation of σ. On the other hand, if we were able to ensure a small prediction error, more assumptions might be required and our nonasymptotic control is likely to involve unknown nuisance parameters that are hard to estimate.

3 Simulations

In this section, we illustrate the performance of our procedure through a Monte-Carlo experiment based on 100 repetitions. The data are generated according to the following model:

$$Y_i = \sum_{l=1}^{k} v_{il}\alpha_l^* + \gamma^* \exp\left(\sum_{l=1}^{k} v_{il}\tau_l^*\right) + W_i, \qquad i = 1, ..., n.$$

The matrix v consists of n rows, which are fixed realizations of i.i.d. draws from the normal distribution $\mathcal{N}\left(\mathbf{0}_k, \Sigma\right)$ where $\Sigma_{jj} = 1$ and $\Sigma_{jj'} = 0.5$ for $j \neq j'$, $j, j' \in \{1, ..., k\}$. The (i, l)th entry of v is denoted by v_{il}; $\alpha^* = \{\alpha_l^*\}_{l=1}^k \in \mathbb{R}^k$, $\tau^* = \{\tau_l^*\}_{l=1}^k \in \mathbb{R}^k$, and $\gamma^* \in \mathbb{R}$ are the unknown coefficients (as a result, $p = 2k+1$). For each of the 100 repetitions, the noise vector W is randomly drawn from $\mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$ with $\sigma = 0.5$.

Let v_i denote the ith row of v and $g\left(v_i;\theta^*\right) = \sum_{l=1}^{k} v_{il}\alpha_l^* + \gamma^* \exp\left(\sum_{l=1}^{k} v_{il}\tau_l^*\right)$. Our hypotheses take the form

$$H_0 : \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g\left(v_i;\theta^*\right)}{\partial v_{il}} \in [0,\, 0.8]\,, \qquad \forall l \in M \subseteq \{1, ..., k\} \tag{41}$$

$$H_1 : \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g\left(v_i;\theta^*\right)}{\partial v_{il}} \notin [0,\, 0.8]\,, \qquad \forall l \in M \subseteq \{1, ..., k\}$$

where

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial g\left(v_i;\theta^*\right)}{\partial v_{il}} = \alpha_l^* + \gamma^*\tau_l^* \frac{1}{n}\sum_{i=1}^{n} \exp\left(\sum_{l=1}^{k} v_{il}\tau_l^*\right)$$

is the APE of v_{il}s on $\mathbb{E}\left(Y_i\right) = g\left(v_i;\theta^*\right)$. Two samples sizes, $n = \{30,\, 90\}$, are considered in our experiment. We first look at a low-dimensional scenario where $k = 1$ (i.e., $v_i = v_{i1}$) and θ^* consists of 3 unknown components ($p = 3$). For this scenario, we have $M = \{1\}$ in (41) ($m = 1$). We then look at a high-dimensional

scenario where $k = 15$ (i.e., $v_i = \{v_{il}\}_{l=1}^{15}$) and θ^* consists of 31 unknown components ($p = 31$). For this scenario, we consider $M = \{1, ..., k\}$ in (41) ($m = 15$). To examine the impact of L, we look at the choice $X_i = (v_{i1},\, v_{i1}^2,\, v_{i1}^3) \in \mathbb{R}^3$ versus the choice $X_i = (v_{i1},\, v_{i1}^2,\, v_{i1}^3,\, v_{i1}^4) \in \mathbb{R}^4$ in the low-dimensional scenario, and $X_i = \left\{v_{ij},\, v_{ij}^2,\, v_{ij}^3\right\}_{j=1}^{15} \in \mathbb{R}^{45}$ versus $X_i = \left\{v_{ij},\, v_{ij}^2,\, v_{ij}^3,\, v_{ij}^4\right\}_{j=1}^{15} \in \mathbb{R}^{60}$ in the high-dimensional scenario. The entries in X are normalized such that $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = 1$ for all $j = 1, ..., L$.

We apply program (4) with $q = \infty$ and set $r_{\alpha,\infty}$ to the minimum of (24) (with $q = \infty$) and (36). For each of the 100 repetitions, we take $R = 10000$ i.i.d. draws (Z_rs) from $\mathcal{N}(\mathbf{0}_n,\, \boldsymbol{I}_n)$ and choose $\alpha_1 = 0.049$, $\alpha_2 = 0.001$ (i.e., $\alpha = 0.05$) to balance between $\tau_{\alpha_1,\infty}$ and $\sqrt{\frac{1}{R}}\tau_{\alpha_2,\infty}$ in (24). For (36), we simply choose $\alpha = 0.05$. We compare the actual separation $\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[g(V_i;\theta^*) - g\left(V_i;\hat{\theta}_\alpha\right)\right]\right\|_\infty$ with $2r_{\alpha,\infty}$. By setting $\beta_1 = 0.001$, $\beta_2 = 0.049$ in (31) (and $\beta = 0.05$ in (37)), note that $2r_{\alpha,\infty}$ approximates (respectively, coincides with) $\delta_{\alpha,\beta,\infty}$. Tables 1 and 2 exhibit:

(i) $\frac{1}{n}\sum_{i=1}^n \frac{\partial g(v_i;\theta^*)}{\partial v_{il}}$ for $l \in M$,

(ii) the average of $\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[g(V_i;\theta^*) - g\left(V_i;\hat{\theta}_\alpha\right)\right]\right\|_\infty$ over 100 repetitions,

(iii) the average of $2r_{\alpha,\infty}$ over 100 repetitions,

(iv) the coverage probability,

(v) the rejection probability.

The evidence from our simulation study supports the main points of Section 2. For our procedure to reject the null hypothesis, all it takes is sufficient separation in $\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[g(V_i;\theta^*) - g\left(V_i;\hat{\theta}_\alpha\right)\right]\right\|_\infty$ and θ^* need not have any form of sparsity. Also in view of (26) and (27), it is not surprising that the coverage probabilities of our procedure are not affected by whether θ^* is sparse or not.

In contrast to "undercoverage" commonly reported in many asymptotic procedures, the coverage probabilities shown in Tables 1 and 2 suggest that our method tends to be conservative. The actual separation (ii) needed to achieve a power of around 95% or higher is somewhat comparable to (and smaller than) the theoretical prediction (iii). This result is plausible given that our control on the Type II error only states that β is an upper bound on the probability of our procedure failing to reject H_0.

As shown in Tables 1 and 2, the values of $\frac{1}{n}\sum_{i=1}^n \frac{\partial g(v_i;\theta^*)}{\partial v_{il}}$s (i) that make most of the rejection probabilities around 95% or higher decrease as n increases from 30 to 90. This finding is intuitive: keeping all the other factors the same, the numbers of constraints and "free" parameters to be determined in (4) remain unchanged but $r_{\alpha,\infty}$ decreases as n increases. Consequently, it takes smaller $\frac{1}{n}\sum_{i=1}^n \frac{\partial g(v_i;\theta^*)}{\partial v_{il}}$s for our procedure to correctly reject H_0. In addition, a larger L improves the power in

most cases, for which we have provided an explanation in Section 2.1.4.

Table 1: $p = 3$, $M = \{1\}$					
	i	**ii**	**iii**	**iv**	**v**
$n = 30$, $L = 4$	1.482	0.091	0.106	1	93%
$n = 30$, $L = 3$	1.482	0.068	0.103	1	64%
$n = 90$, $L = 4$	1.159	0.028	0.035	1	93%
$n = 90$, $L = 3$	1.159	0.028	0.034	1	96%

For $n = 30$, $\alpha^* = \tau^* = 0.657$, $\gamma^* = 1$
For $n = 90$, $\alpha^* = \tau^* = 0.533$, $\gamma^* = 1$

Table 2: $p = 31$, $M = \{1, ..., 15\}$					
	i	**ii**	**iii**	**iv**	**v**
$n = 30$, $L = 60$	$(1.735, ..., 1.735)$	0.099	0.132	1	100%
$n = 30$, $L = 45$	$(1.735, ..., 1.735)$	0.096	0.129	1	95%
$n = 90$, $L = 60$	$(0.929, ..., 0.929)$	0.035	0.044	1	99%
$n = 90$, $L = 45$	$(0.929, ..., 0.929)$	0.033	0.043	1	94%

For $n = 30$, $\alpha^* = \tau^* = (0.194, ..., 0.194)$, $\gamma^* = 1$
For $n = 90$, $\alpha^* = \tau^* = (0.172, ..., 0.172)$, $\gamma^* = 1$

4 Extensions

Beyond Gaussian regressions, it is possible to establish some nonasymptotic justifications for inference in statistical models that involve non-Gaussian responses.

4.1 Regressions with non-Gaussian noise

Our analyses in Section 2 exploit sharp concentration of Lipschitz functions of Gaussian variables. These analyses can be extended to regression models where the noise vector W is either bounded or has a strongly log-concave distribution. In particular, we have the following analogues of (16).

Lemma 4.1. *Suppose W has a strongly log-concave[2] distribution with parameter*

[2]A strongly log-concave distribution is a distribution with density $\mathsf{p}(z) = \exp(-\psi(z))$ such that for some $\varphi > 0$ and all $\lambda \in [0, 1]$, z, $z^{'} \in \mathbb{R}^n$, $\lambda\psi(z) + (1-\lambda)\psi(z^{'}) - \psi(\lambda z + (1-\lambda)z^{'}) \geq \frac{\varphi}{2}\lambda(1-\lambda)\left\|z - z^{'}\right\|_2^2$.

φ. *Then for any* $q \in [1, \infty]$, *we have*

$$\mathbb{P}\left\{\left\|\frac{1}{n}X^T W\right\|_q \geq \mathbb{E}\left[\left\|\frac{1}{n}X^T W\right\|_q\right] + t\right\} \leq \exp\left(\frac{-n\varphi t^2}{2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q^2}\right). \tag{42}$$

Remarks. Let $g(V;\theta^*) := \{g(V_i;\theta^*)\}_{i=1}^n$. For a fixed design V, if $Y \sim \mathcal{N}(g(V;\theta^*), \Sigma)$ and $\Sigma \succ 0$, φ can be set to the smallest eigenvalue of Σ^{-1}. Beyond a normal distribution, [17] discuss quite a few examples of strongly log-concave distributions.

Lemma 4.2. *Suppose* W *consists of independent random variables, all of which are supported on* $[a, b]$. *Then for any* $q \in [1, \infty]$, *we have*

$$\mathbb{P}\left\{\left\|\frac{1}{n}X^T W\right\|_q \geq \mathbb{E}\left[\left\|\frac{1}{n}X^T W\right\|_q\right] + t\right\} \leq \exp\left(\frac{-nt^2}{2(b-a)^2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q^2}\right). \tag{43}$$

If we know the distribution of W, our analyses from Section 2 can be, in principle, extended to construct inference procedures for regression models where W is either bounded or has a strongly log-concave distribution. However, sometimes we might not know the distribution for W; instead, we may have more information on the distribution of Y than the distribution of W. In some applications, we might know Y consists of entries supported on $[a, b]$. For example, [16] and [22] estimate the APE of spending on math pass rates ($Y_i \in [0, 1]$) under the assumption $\mathbb{E}(Y_i|V_i) = \Phi(V_i\theta^*)$, where $\Phi(\cdot)$ denotes the standard normal c.d.f. and V_i include the spending variable as well as other covariates. Another example is the binary response model

$$\mathbb{P}(Y_i = 1|V_i) = g(V_i;\theta^*), \qquad i = 1, ..., n, \tag{44}$$

where $Y_i \in \{0, 1\}$ and the functional form of $g(V_i;\theta^*)$ is known; for example, g may be a "probit" or a "logit" in (44) and $g(V_i;\theta^*) = g(V_i\theta^*)$. Under the assumption

$$\mathbb{E}(Y_i|V_i) = g(V_i;\theta^*), \tag{45}$$

both binary and bounded response models can be treated in the same framework.

4.2 Bounded responses

In what follows, we consider (45) where $a \leq Y_i \leq b$ for all i, the functional form of $g(V_i;\theta^*)$ is known and possibly nonlinear in θ^*. Without loss of generality, we assume $a = 0$ and $b = 1$. Throughout this section, we use $\mathbb{E}_{Y|V}[\cdot]$ to denote the expectation over the distribution of Y conditioning on V; for an i.i.d. sequence of Radamacher random variables, $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ (independent of Y and V), we use $\mathbb{E}_\varepsilon[\cdot]$ to denote the expectation over ε only, conditioning on Y and V, and $\mathbb{E}_{\varepsilon,Y|V}[\cdot]$ to denote the expectation over the distribution of (ε, Y) conditioning on V.

As in Section 2.2, for $q = \infty$, the choice of $r_{\alpha,\infty}$ can be based on simple union bounds. We omit discussions of this strategy but focus on the concentration approach for the more general cases $q \in [1, \infty]$. Like in the regression problem, we first establish the concentration of

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - g\left(V_i; \theta^*\right) \right] \right\|_q$$

around its expectation

$$S_{\theta^*} := \mathbb{E}_{Y|V} \left[\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - g\left(V_i; \theta^*\right) \right] \right\|_q \right]. \tag{46}$$

Previously we have simply replaced $\mathbb{E}_W \left[\left\| \frac{1}{n} X^T W \right\|_q \right]$ in (16) with its Monte Carlo approximation $\frac{\sigma}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} X^T Z_r \right\|_q$ and a "small" deviation. This strategy cannot be applied to the expectation S_{θ^*} directly. Instead, we first seek a reasonable upper bound which involves only $\{Y, X\}$ and random variables from a known distribution. This result is stated in the following proposition.

Proposition 4.1. *Assume $Y = \{Y_i\}_{i=1}^n$ consists of independent random variables, where $0 \leq Y_i \leq 1$ for all i (w.l.o.g.). For any $q \in [1, \infty]$, we have*

$$\mathbb{P}\left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - g\left(V_i; \theta^*\right) \right] \right\|_q \geq S_{\theta^*} + t \right\} \leq \exp\left(\frac{-nt^2}{2 \left\| \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_i^2} \right\|_q^2} \right). \tag{47}$$

Let $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ be an i.i.d. sequence of Radamacher random variables independent of Y and V. Under (45), we have

$$\mathbb{E}_{\varepsilon,Y|V} \left\{ \left\| \frac{1}{2n} \sum_{i=1}^{n} \varepsilon_i X_i \left[Y_i - g\left(V_i; \theta^*\right) \right] \right\|_q \right\} \leq S_{\theta^*} \leq 2\mathbb{E}_{\varepsilon,Y|V} \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i Y_i \right\|_q \right\}. \tag{48}$$

Remarks. Note that bound (47) holds for any fixed θ (not just the true coefficient vector, θ^*). However, (48) relies crucially on the model assumption (45).

The upper bound in (48) can be viewed as the symmetrized version of S_{θ^*}. Considering a collection of i.i.d. Radamacher random draws (independent of Y and V),

$$\{\varepsilon_{ir} : i = 1, ..., n, r = 1, ..., R\}, \tag{49}$$

we can replace S_{θ^*} with $\frac{2}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ir} Y_i X_i \right\|_q$ (a Monte-Carlo approximation of the symmetrized version) and some "small" deviations. The complementary lower bound in (48) suggests that S_{θ^*} and its symmetrized version have the same magnitude.

The first step is to relate S_{θ^*} with $\frac{2}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q$ as shown in the following proposition.

Proposition 4.2. *Assume (45) where $Y=\{Y_i\}_{i=1}^n$ consists of independent random variables, $0\le Y_i\le 1$ for all i (w.l.o.g.). Given (49) which is independent of Y and V, for any $q\in[1,\infty]$, we have*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left[Y_i-g\left(V_i;\theta^*\right)\right]\right\|_q \ge \frac{2}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q + t_1+2t_2+2t_3 \tag{50}$$

with probability no greater than $\alpha\in(0,1)$, where

$$\begin{aligned} t_1 &= \tau_{\alpha_1,q}=\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\sqrt{\frac{2}{n}\log\frac{1}{\alpha_1}},\\ t_2 &= \tau_{\alpha_2,q}=\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\sqrt{\frac{2}{n}\log\frac{1}{\alpha_2}},\\ t_3 &= \frac{2}{\sqrt{R}}\tau_{\alpha_3,q}=\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\sqrt{\frac{8}{nR}\log\frac{1}{\alpha_3}}, \end{aligned}$$

for chosen α_1, α_2, $\alpha_3>0$ such that $\sum_{k=1}^{3}\alpha_k=\alpha$.

Based on (50) along with the choices of t_1, t_2 and t_3 above, we set in (4),

$$r_{\alpha,q}=\frac{2}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q+\tau_{\alpha_1,q}+2\tau_{\alpha_2,q}+\frac{4}{\sqrt{R}}\tau_{\alpha_3,q}. \tag{51}$$

Under H_0, $(\theta^*,0)$ is an optimal solution to (4) with $r_{\alpha,q}$ specified in (51). Consequently, a (practical) optimal solution to (4) must satisfy

$$\mathbb{P}_0\left\{\Psi_q(\hat{\theta}_\alpha)\ge r_{\alpha,q}\right\}\le\alpha. \tag{52}$$

Moreover, following the argument used for deriving (18) and (19), we can show that, an optimal solution $\left(\hat{\theta}_\alpha,\hat{\mu}_\alpha\right)$ to (4) must satisfy

$$\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left[g\left(V_i;\theta^*\right)-g\left(V_i;\hat{\theta}_\alpha\right)\right]\right\|_q \ge \hat{\mu}_\alpha, \tag{53}$$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left[g\left(V_i;\theta^*\right)-g\left(V_i;\hat{\theta}_\alpha\right)\right]\right\|_q \le 2r_{\alpha,q}+\hat{\mu}_\alpha, \tag{54}$$

with probability at least $1-\alpha$.

For general $q \in [1, \infty]$, the strategy where we replace S_{θ^*} in (46) by

$$\frac{2}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q$$

plus some deviations only requires the correct specification of the conditional mean of Y_i, i.e., (45). This treatment delivers generic confidence regions in the form of (53) and (54) for binary and fractional responses.

Acknowledgment

I am grateful to the associate editor's insightful comments. I also thank Guang Cheng at Purdue University for discussions that improved the clarity of an earlier version (arXiv:1808.07127v1) and pointing out several references. All errors are my own.

A Supplementary materials

A.1 Alternative formulation

Instead of (4), we can work with an alternative formulation:

$$\begin{aligned}\left(\hat{\theta}_\alpha,\,\hat{\mu}_\alpha\right) \in \arg\min_{(\theta_\alpha,\mu_\alpha)\in\mathbb{R}^p\times\mathbb{R}^p} \|\mu_\alpha\|_{\tilde{q}} \\ \text{subject to: } \left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left[Y_i - g\left(V_i;\theta_\alpha\right)\right] - \mu_\alpha\right\|_q \le r_{\alpha,q}, \\ h(\theta_\alpha)\in\Omega.\end{aligned} \tag{55}$$

Here we slightly abuse the notations, where $\hat{\mu}_\alpha$ (also μ_α) in (55) is a vector and in (4) is a scalar. Like in (4), we suppress the dependence of $(\hat{\theta}_\alpha, \hat{\mu}_\alpha)$ in (55) on $(q,\, \tilde{q})$ for notational simplicity.

To compare (55) with (4) from the computational perspective, we let $\mathcal{F}_1^\alpha$ denote the set of $(\theta_\alpha, \mu_\alpha)$ that are feasible for (55) and $\mathcal{F}_{1,\theta}^\alpha$ denote the set of θ_α from $\mathcal{F}_1^\alpha$; similarly, $\mathcal{F}_2^\alpha$ and $\mathcal{F}_{2,\theta}^\alpha$ are defined with regard to (4). Note that an element $(\tilde{\theta}_\alpha, \tilde{\mu}_\alpha)$ in $\mathcal{F}_1^\alpha$ implies

$$\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left[Y_i - g\left(V_i;\tilde{\theta}_\alpha\right)\right]\right\|_q \le r_{\alpha,q} + \|\tilde{\mu}_\alpha\|_q\,;$$

that is, $(\tilde{\theta}_\alpha, \|\tilde{\mu}_\alpha\|_q) \in \mathcal{F}_2^\alpha$. Consequently, $\mathcal{F}_{1,\theta}^\alpha \subseteq \mathcal{F}_{2,\theta}^\alpha$. On the other hand, the objective function in (55) is minimized over an $L-$dimensional vector as opposed to a scalar in (4). These facts suggest that the choice between (55) and (4) incurs some trade-offs in terms of computational cost.

Theorem 2.1 holds for the alternative formulation. Moreover, the following results exhibit the "ideal" and practical confidence regions.

A.1.1 Ideal confidence regions

Proposition A.2.1. *Assume (2) where* $W \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$ *and is independent of* V. *Then, an optimal solution* $\left(\hat{\theta}^*_\alpha, \hat{\mu}^*_\alpha\right)$ *to (55) must satisfy*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \hat{\theta}^*_\alpha\right) \right] \right\|_{\tilde{q}} \geq \|\hat{\mu}^*_\alpha\|_{\tilde{q}}, \tag{56}$$

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \hat{\theta}^*_\alpha\right) \right] - \hat{\mu}^*_\alpha \right\|_{q} \leq 2r^*_{\alpha,q}, \tag{57}$$

with probability at least $1-\alpha$, *where* $r^*_{\alpha,q}$ *is specified in (17).*

A.1.2 Practical confidence regions

Let $\left(\hat{\theta}_\alpha, \hat{\mu}_\alpha\right)$ be an optimal solution to (55) with $r_{\alpha,q}$ specified in (24). Then we have

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \hat{\theta}_\alpha\right) \right] - \hat{\mu}_\alpha \right\|_{q} \\ \leq & \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[Y_i - g\left(V_i; \hat{\theta}_\alpha\right) \right] - \hat{\mu}_\alpha \right\|_{q} + \left\| \frac{1}{n} X^T W \right\|_{q} \\ \leq & \frac{2\sigma}{R} \sum_{r=1}^{R} \left\| \frac{1}{n} X^T Z_r \right\|_{q} + 2\tau_{\alpha_1,q} + 2\sqrt{\frac{1}{R}} \tau_{\alpha_2,q} \end{aligned} \tag{58}$$

and

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \left[g\left(V_i; \theta^*\right) - g\left(V_i; \hat{\theta}_\alpha\right) \right] \right\|_{\tilde{q}} \geq \|\hat{\mu}_\alpha\|_{\tilde{q}} \tag{59}$$

with probability at least $1-\alpha$.

A.2 Preliminary

Here we include several classical results which are used in the main proofs. We first introduce a definition of sub-Gaussian variables.

Definition A.1. *A zero-mean random variable* U_1 *is sub-Gaussian if there is a* $\nu > 0$ *such that*

$$\mathbb{E}\left[\exp\left(\lambda U_1\right)\right] \leq \exp\left(\frac{\lambda^2 \nu^2}{2}\right) \tag{60}$$

for all $\lambda \in \mathbb{R}$, *and we refer to* ν *as the sub-Gaussian parameter.*

Remarks.

1. Using the Chernoff bound, one can show that any zero-mean random variable U_1 obeying (60) satisfies

$$\mathbb{P}\left(U_1 \leq -t\right) \leq \exp\left(-\frac{t^2}{2\nu^2}\right), \tag{61}$$

$$\mathbb{P}\left(U_1 \geq t\right) \leq \exp\left(-\frac{t^2}{2\nu^2}\right), \tag{62}$$

for all $t \geq 0$.

2. Let $\{U_i\}_{i=1}^R$ be independent zero-mean sub-Gaussian random variables, each with parameter at most ν. Then $R^{-1}\sum_{i=1}^R U_i$ is sub-Gaussian with parameter at $\nu/\sqrt{R}$. To see this, note that for all $\lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\lambda}{R}\sum_{i=1}^R U_i\right)\right] &= \prod_{i=1}^R \mathbb{E}\left[\exp\left(\frac{\lambda U_i}{R}\right)\right] \\ &\leq \prod_{i=1}^R \exp\left(\frac{\lambda^2\nu^2}{2R^2}\right) \\ &= \exp\left(\frac{\lambda^2\nu^2}{2R}\right). \end{aligned} \tag{63}$$

The following result exhibits the type of sub-Gaussian variables that are of interest to our analysis.

Lemma A.1. *Suppose $U = \{U_i\}_{i=1}^n$ has a strongly log-concave distribution with parameter $\varphi > 0$ and $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{L}-$Lipschitz with respect to the Euclidean norm. Then for all $\lambda \in \mathbb{R}$, we have*

$$\mathbb{E}\left[\exp\left(\lambda\left\{f(U) - \mathbb{E}\left[f(U)\right]\right\}\right)\right] \leq \exp\left(\frac{\lambda^2\mathcal{L}^2}{2\varphi}\right). \tag{64}$$

As a consequence,

$$\begin{aligned}\mathbb{P}\left\{f(U) - \mathbb{E}\left[f(U)\right] \leq -t\right\} &\leq \exp\left(-\frac{\varphi t^2}{2\mathcal{L}^2}\right), \\ \mathbb{P}\left\{f(U) - \mathbb{E}\left[f(U)\right] \geq t\right\} &\leq \exp\left(-\frac{\varphi t^2}{2\mathcal{L}^2}\right).\end{aligned}$$

Remarks. The proof involves the so-called "inf-convolution" argument and an application of the Brunn-Minkowski inequality; see [6] and [14].

Lemma A.2. *Assume $U = \{U_i\}_{i=1}^n$ consists of independent random variables,*

all of which are supported on $[a, b]$. *If* $f : \mathbb{R}^n \to \mathbb{R}$ *is separately convex*[3] *and* $\mathcal{L}-$*Lipschitz with respect to the Euclidean norm, then for all* $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda\left\{f(U) - \mathbb{E}\left[f(U)\right]\right\}\right)\right] \leq \exp\left[\frac{\lambda^2(b-a)^2\mathcal{L}^2}{2}\right]. \tag{65}$$

As a consequence,

$$\begin{aligned}
\mathbb{P}\left[f(U) - \mathbb{E}\left[f(U)\right] \leq -t\right] &\leq \exp\left(-\frac{t^2}{2\mathcal{L}^2(b-a)^2}\right), \\
\mathbb{P}\left[f(U) - \mathbb{E}\left[f(U)\right] \geq t\right] &\leq \exp\left(-\frac{t^2}{2\mathcal{L}^2(b-a)^2}\right).
\end{aligned}$$

Remarks. One proof for Lemma A.2 involves the entropy method and the so-called Herbst argument; see [7]. Talagrand and Ledoux have contributed to the result above in different papers.

A.3 Proof of Propositions 2.1 and A.2.1

For any $q \in [1, \infty]$, $\left\|\frac{1}{n}X^TW\right\|_q$ is Lipschitz in W with respect to the Euclidean norm. To see this, note that a triangle inequality and a Cauchy-Schwarz inequality yield

$$\begin{aligned}
\left|\left\|\frac{1}{n}X^TW\right\|_q - \left\|\frac{1}{n}X^TW^{'}\right\|_q\right| &\leq \left\|\frac{1}{n}X^T\left(W - W^{'}\right)\right\|_q \\
&\leq \frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q \left\|W - W^{'}\right\|_2.
\end{aligned} \tag{66}$$

As a result of Lemma A.1, we have the concentration in (16).

If $h(\theta^*) \in \Omega$, (16) then implies that $(\theta^*, \mathbf{0}_p)$ $((\theta^*, 0))$ is an optimal solution to (55) (respectively, (4)). If $h(\theta^*) \notin \Omega$, since $\{\theta \in \mathbb{R}^p : h(\theta) \in \Omega\} \neq \emptyset$, we can find some $\tilde{\theta}_\alpha$ such that $h(\tilde{\theta}_\alpha) \in \Omega$. Letting

$$\begin{aligned}
\tilde{\mu}_\alpha &= \frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - g\left(V_i; \tilde{\theta}_\alpha\right)\right] - \frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - g\left(V_i; \theta^*\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^n X_i\left[g\left(V_i; \theta^*\right) - g\left(V_i; \tilde{\theta}_\alpha\right)\right],
\end{aligned}$$

[3]Let the function $f_j : \mathbb{R} \to \mathbb{R}$ be defined by varying only the jth co-ordinate of a function $f : \mathbb{R}^n \to \mathbb{R}$; f is *separately convex* if for each $j \in \{1, 2, ..., n\}$, f_j is a convex function of the jth coordinate.

(16) then implies that $\left(\tilde{\theta}_\alpha,\, \tilde{\mu}_\alpha\right)$ is a feasible solution to (55) with probability at least $1-\alpha$. As a result, $\left(\tilde{\theta}_\alpha,\, \|\tilde{\mu}_\alpha\|_q\right)$ is also a feasible solution to (4) with probability at least $1-\alpha$.

In any case, an optimal solution $\left(\hat{\theta}^*_\alpha,\, \hat{\mu}^*_\alpha\right)$ to (55) must satisfy

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - g\left(V_i;\hat{\theta}^*_\alpha\right)\right] - \frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - g\left(V_i;\theta^*\right)\right]\right\|_{\tilde{q}} \\
=&\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[g\left(V_i;\theta^*\right) - g\left(V_i;\hat{\theta}^*_\alpha\right)\right]\right\|_{\tilde{q}} \geq \|\hat{\mu}^*_\alpha\|_{\tilde{q}}
\end{aligned}$$

with probability at least $1-\alpha$. Similarly, an optimal solution $\left(\hat{\theta}^*_\alpha,\, \hat{\mu}^*_\alpha\right)$ to (4) must satisfy

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - g\left(V_i;\hat{\theta}^*_\alpha\right)\right] - \frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - g\left(V_i;\theta^*\right)\right]\right\|_{q} \\
=&\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[g\left(V_i;\theta^*\right) - g\left(V_i;\hat{\theta}^*_\alpha\right)\right]\right\|_{q} \geq \hat{\mu}^*_\alpha
\end{aligned}$$

with probability at least $1-\alpha$.

On the other hand, in terms of (55), applying the triangle inequality yields

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[g\left(V_i;\theta^*\right) - g\left(V_i;\hat{\theta}^*_\alpha\right)\right] - \hat{\mu}^*_\alpha\right\|_q \\
\leq&\left\|\frac{1}{n}X^TW\right\|_q + \left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - g\left(V_i;\hat{\theta}^*_\alpha\right)\right] - \hat{\mu}^*_\alpha\right\|_q \leq 2r^*_{\alpha,q}
\end{aligned}$$

with probability at least $1-\alpha$. In terms of (4), we simply have

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[g\left(V_i;\theta^*\right) - g\left(V_i;\hat{\theta}^*_\alpha\right)\right]\right\|_q \leq 2r^*_{\alpha,q} + \hat{\mu}^*_\alpha\right) \geq 1-\alpha.$$

A.4 Proof of Theorem 2.1

We have already derived (32) in Section 2. To show (33), we define the event

$$\mathcal{E} = \left\{\frac{\sigma}{R}\sum_{r=1}^R\left\|\frac{1}{n}X^TZ_r\right\|_q \geq \mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right] + \sqrt{\frac{1}{R}}\tau_{\beta_1,q}\right\}.$$

Like we have argued for (21), we also have the upper deviation inequality

$$\mathbb{P}\left\{\frac{\sigma}{R}\sum_{r=1}^R\left\|\frac{1}{n}X^TZ_r\right\|_q \geq \mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right] + t\right\} \leq \exp\left(\frac{-nRt^2}{2\sigma^2\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q^2}\right) \tag{67}$$

and consequently, $\mathbb{P}(\mathcal{E}) \leq \beta_1$. Let $\mathcal{E}^c$ denote the complement of $\mathcal{E}$. Under H_1 and (30), we have

$$\begin{aligned}
&\mathbb{P}\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q}\right\} \\
=&\mathbb{P}\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q}|\mathcal{E}^c\right\}\mathbb{P}(\mathcal{E}^c) + \mathbb{P}\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q}|\mathcal{E}\right\}\mathbb{P}(\mathcal{E}) \\
\leq&\mathbb{P}\left\{\Psi_q(\hat{\theta}_\alpha) \leq r_{\alpha,q}|\mathcal{E}^c\right\} + \mathbb{P}(\mathcal{E}) \\
\leq&\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[g(V_i;\theta^*) - g\left(V_i;\hat{\theta}_\alpha\right)\right]\right\|_q - \left\|\frac{1}{n}X^TW\right\|_q \leq r_{\alpha,q}|\mathcal{E}^c\right\} + \beta_1 \\
\leq&\mathbb{P}\left\{\delta_{\alpha,\beta,q} - \left\|\frac{1}{n}X^TW\right\|_q \leq r_{\alpha,q}|\mathcal{E}^c\right\} + \beta_1 \\
\leq&\mathbb{P}\left\{\left\|\frac{1}{n}X^TW\right\|_q \geq \mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_q\right] + \tau_{\beta_2,q}|\mathcal{E}^c\right\} + \beta_1 \\
\leq&\beta.
\end{aligned}$$

In the above, the fifth line follows from (30); the sixth line follows from (31), the fact that we are conditioning on $\mathcal{E}^c$, as well as (16); the last line follows from the fact that W is independent of Z.

A.5 Additional derivations

To show (39), we define an i.i.d. sequence of Gaussian random variables

$$\widetilde{W}_k \sim \mathcal{N}\left(0, \min_{j,l\in\{1,...,L\}} \frac{1}{2n^2}\sum_{i=1}^n (X_{ij} - X_{il})^2\right)$$

for $k = 1, ..., L$. Note that we have

$$\mathbb{E}_W\left[\left(\frac{1}{n}X_j^TW - \frac{1}{n}X_l^TW\right)^2\right] \geq \mathbb{E}_{\widetilde{W}}\left[\left(\widetilde{W}_j - \widetilde{W}_l\right)^2\right].$$

By the Sudakov-Fernique Gaussian comparison result (see Corollary 3.14 in [13]), we obtain

$$\begin{aligned}
\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] &\geq \mathbb{E}_W\left[\max_{j\in\{1,...,L\}}\frac{1}{n}X_j^TW\right] \\
&\geq \frac{1}{2}\mathbb{E}_{\widetilde{W}}\left[\max_{j\in\{1,...,L\}}\widetilde{W}_j\right] \\
&\geq \frac{1}{2}\left(1 - \frac{1}{e}\right)\sqrt{\frac{\log L}{4n^2}\min_{j,l\in\{1,...,L\}}\sum_{i=1}^n (X_{ij} - X_{il})^2}
\end{aligned}$$

(for all $L \geq 20$), where the last line follows from a classical lower bound on the Gaussian maximum (see, e.g., [13]). The upper bound

$$\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right] \leq \sqrt{\frac{2\log L}{n^2}\max_{j\in\{1,\ldots,L\}}\sum_{i=1}^n X_{ij}^2} + \sqrt{\frac{8}{n^2\log L}\max_{j\in\{1,\ldots,L\}}\sum_{i=1}^n X_{ij}^2}$$

(for all $L \geq 2$) is another existing result on the Gaussian maximum (see, e.g., [20]).

Remarks. To obtain the lower bound on $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right]$, we first compare the dependent sequence $\left\{\frac{1}{n}X_j^TW\right\}_{j=1}^L$ with another independent Gaussian sequence $\widetilde{W} = \left\{\widetilde{W}_j\right\}_{j=1}^L$ and then apply a lower bound on $\mathbb{E}_{\widetilde{W}}\left[\max_{j\in\{1,\ldots,L\}}\widetilde{W}_j\right]$. In contrast, the upper bound on $\mathbb{E}_W\left[\left\|\frac{1}{n}X^TW\right\|_\infty\right]$ is obtained by applying $\sum_{j=1}^L \mathbb{P}\left(\left|\frac{1}{n}X_j^TW\right| \geq t\right)$, where independence is not needed. Moreover, the upper bound also holds when W is a sequence of sub-Gaussian variables while the lower bound requires W to be a sequence of Gaussian variables.

A.6 Proofs of Lemmas 4.1 and 4.2

As a result of Lemma A.1 and (66), we have the concentration in Lemma 4.1. Because $\left\|\frac{1}{n}X^TW\right\|_q$ is separately convex in terms of W, Lemma A.2 implies the concentration in Lemma 4.2.

A.7 Proof of Proposition 4.1

Using the argument that leads to (66), we can show $\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - g\left(V_i;\theta^*\right)\right]\right\|_q$ is Lipschitz in Y with respect to the Euclidean norm for any $q \in [1,\infty]$. That is,

$$\begin{aligned}&\left|\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - g\left(V_i;\theta^*\right)\right]\right\|_q - \left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i^{'} - g\left(V_i;\theta^*\right)\right]\right\|_q\right| \\ \leq &\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}\right\|_q \left\|Y - Y^{'}\right\|_2. \end{aligned} \tag{68}$$

Note that $\left\|\frac{1}{n}\sum_{i=1}^n X_i\left[Y_i - g\left(V_i;\theta^*\right)\right]\right\|_q$ is separately convex in terms of Y. As a result of Lemma A.2, we have the concentration in (47).

To establish (48), we exploit the convexity of l_q-norms and the fact that $\mathbb{E}\left(Y_i|V_i\right) = g\left(V_i;\theta^*\right)$. Let $Y^{'} = \left\{Y_i^{'}\right\}_{i=1}^n$ be an independent sequence identical to but independent of Y conditioning on V, and $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ be i.i.d. Radamacher random variables

independent of Y, $Y^{'}$, and V. We obtain

$$\begin{aligned}
&\mathbb{E}_{Y|V}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left[Y_i-g\left(V_i;\theta^*\right)\right]\right\|_q\right\}\\
=&\mathbb{E}_{Y|V}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left[Y_i-\mathbb{E}_{Y_i^{'}|V_i}\left(Y_i^{'}\right)\right]\right\|_q\right\}\\
=&\mathbb{E}_{Y|V}\left\{\left\|\mathbb{E}_{Y^{'}|V}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\left(Y_i-Y_i^{'}\right)\right]\right\|_q\right\}\\
\leq&\mathbb{E}_{Y^{'},Y|V}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left(Y_i-Y_i^{'}\right)\right\|_q\right\}\\
=&\mathbb{E}_{\varepsilon,Y^{'},Y|V}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_i\left(Y_i-Y_i^{'}\right)\right\|_q\right\}\\
\leq&2\mathbb{E}_{\varepsilon,Y|V}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right\},
\end{aligned}\tag{69}$$

where the second line follows since $\mathbb{E}\left(Y_i^{'}|V_i\right)=g\left(V_i;\theta^*\right)$, the fourth line follows from Jensen's inequality, and the sixth line follows from the fact that $\varepsilon_iX_i\left(Y_i-Y_i^{'}\right)$ and $X_i\left(Y_i-Y_i^{'}\right)$ have the same distribution.

On the other hand, similar argument from above also yields

$$\begin{aligned}
&\mathbb{E}_{\varepsilon,Y|V}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}\varepsilon_iX_i\left[Y_i-g\left(V_i;\theta^*\right)\right]\right\|_q\right\}\\
=&\mathbb{E}_{\varepsilon,Y|V}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}\varepsilon_iX_i\left[Y_i-\mathbb{E}_{Y_i^{'}|V_i}\left(Y_i^{'}\right)\right]\right\|_q\right\}\\
\leq&\mathbb{E}_{\varepsilon,Y^{'},Y|V}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}\varepsilon_iX_i\left(Y_i-Y_i^{'}\right)\right\|_q\right\}\\
=&\mathbb{E}_{Y^{'},Y|V}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i-Y_i^{'}\right)\right\|_q\right\}.
\end{aligned}$$

Applying the following inequality

$$\begin{aligned}
&\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i-Y_i^{'}\right)\right\|_q\\
\leq&\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i-g\left(V_i;\theta^*\right)\right)\right\|_q+\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i^{'}-g\left(V_i;\theta^*\right)\right)\right\|_q,
\end{aligned}$$

and taking expectations gives

$$\mathbb{E}_{Y',Y|V}\left\{\left\|\frac{1}{2n}\sum_{i=1}^{n}X_i\left(Y_i-Y_i^{'}\right)\right\|_q\right\}\leq\mathbb{E}_{Y|V}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\left(Y_i-g\left(V_i;\theta^*\right)\right)\right\|_q\right\}.$$

Putting the pieces together, we obtain the result in (48).

A.8 Proof of Proposition 4.2

We first show that $\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right\}$ is Lipschitz in Y with respect to the Euclidean norm for any $q\in[1,\,\infty]$. That is,

$$\begin{aligned}&\left|\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right\}-\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i^{'}\right\|_q\right\}\right|\\ \leq&\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\sqrt{\mathbb{E}_\varepsilon\left[\sum_{i=1}^{n}\varepsilon_i^2\left(Y_i-Y_i^{'}\right)^2\right]}\\ \leq&\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q\left\|Y-Y^{'}\right\|_2.\end{aligned}$$

Note that $\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right\}$ is separately convex in terms of Y. As a result of Lemma A.2, we have the following concentration

$$\mathbb{P}\left\{\mathbb{E}_{\varepsilon,Y|V}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right\}\geq\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right\}+\tau_{\alpha_2,q}\right\}\leq\alpha_2. \tag{70}$$

Let $\varepsilon=\{\varepsilon_i\}_{i=1}^n$ be an i.i.d. sequence of Radamacher random variables, independent of Y and V. We can again show that $\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iY_iX_i\right\|_q$ is Lipschitz in ε with respect to the Euclidean norm for any $q\in[1,\,\infty]$ and the Lipschitz constant[4] is $\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_iX_i\right)^2}\right\|_q$, which is bounded from above by $\frac{1}{\sqrt{n}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q$ given

[4]Like $\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q$, we define

$$\begin{aligned}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_iX_i)^2}\right\|_q&=\sqrt[q]{\sum_{j=1}^{L}\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2X_{ij}^2}\right)^q},\qquad q\in[1,\,\infty)\\ \left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_iX_i)^2}\right\|_q&=\max_{j\in\{1,\ldots,L\}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2X_{ij}^2},\qquad q=\infty.\end{aligned}$$

$0 \leq Y_i \leq 1$. Let $\{\varepsilon_{ir} : i = 1, ..., n,\ r = 1, ..., R\}$ be a collection of i.i.d. Radamacher random draws, independent of Y and V. Conditioning on Y and V, (63) and (65) imply $\frac{1}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q - \mathbb{E}_\varepsilon\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right)$ is sub-Gaussian with parameter at most $\frac{2}{\sqrt{nR}}\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q$. Therefore, we have

$$\mathbb{E}_{Y|V}\left[\mathbb{E}_\varepsilon\left[\exp\left(\lambda\left[\frac{1}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q - \mathbb{E}_\varepsilon\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right)\right]\right)\right]\right]$$
$$\leq \exp\left[\lambda^2\frac{4\left\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}\right\|_q^2}{2nR}\right].$$

Consequently, (61) yields the following concentration

$$\mathbb{P}\left\{\mathbb{E}_\varepsilon\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_iY_i\right\|_q\right\} \geq \frac{1}{R}\sum_{r=1}^{R}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ir}Y_iX_i\right\|_q + \frac{2}{\sqrt{R}}\tau_{\alpha_3,q}\right\} \leq \alpha_3 \qquad (71)$$

Combining (47), (69), (70) and (71) yields (50).

References

[1] Arlot, S., G. Blanchard, and E. Roquain (2010). "Some Nonasymptotic Results on Resampling in High Dimension, I: Confidence Regions." *Annals of Statistics*, 38, 51-82.

[2] Bertsimas, D. and J. Tsitsiklis (1997). *Introduction to Linear Optimization*, Athena Scientific.

[3] Ben-Tal, A. and A. Nemirovski (2000). "Robust Solutions of Linear Programming Problems Contaminated with Uncertain Data." *Mathematical Programming*, 88, 411-424.

[4] Bertsimas, D. and M. Sim (2004). "The Price of Robustness." *Operations Research*, 52, 35-53.

[5] Bickel, P., J. Y. Ritov, and A. B. Tsybakov (2009). "Simultaneous Analysis of Lasso and Dantzig Selector." *Annals of Statistics*, 37, 1705-1732.

[6] Bobkov, S. G. and M. Ledoux (2000). "From Brunn-Minkowski to Brascamp-Lieb and to Logarithmic Sobolev Inequalities." *Geometric and Functional Analysis*. 10, 1028-1052.

[7] Boucheron, S, G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. Oxford.

[8] Chernozhukov, V., D. Chetverikov, and K. Kato (2013). "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors." *Annals of Statistics*, 41, 2786-2819.

[9] Dezeure, R., P. Bühlmann, and C.-H. Zhang (2017). "High-Dimensional Simultaneous Inference with the Bootstrap." *Test*, 26, 685-719.

[10] Gautier, E. and A. B. Tsybakov (2011). "High-Dimensional Instrumental Variables Regression and Confidence Sets." Manuscript. CREST (ENSAE).

[11] Horowitz, J. L. (2017). "Non-Asymptotic Inference in Instrumental Variables Estimation." Manuscript. Northwestern University.

[12] Javanmard, A. and A. Montanari (2014). "Confidence Intervals and Hypothesis Testing for High- Dimensional Regression." *Journal of Machine Learning Research*, 15, 2869-2909.

[13] Ledoux, M., and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY.

[14] Maurey, B. (1991). "Some Deviation Inequalities." *Geometric and Functional Analysis*. 1, 188-197.

[15] Ning, Y and H. Liu (2017). "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models." *Annals of Statistics*, 45, 158-195.

[16] Papke, L. E. and J. M. Wooldridge (2008). "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates." *Journal of Econometrics*, 145, 121-133.

[17] Saumard, A. and J. A. Wellner (2014). "Log-Concavity and Strong Log-Concavity: A Review." *Statistics Surveys,* 8, 45-114.

[18] Soyster, A. L. (1973). "Convex Programming with Set-Inclusive Constraints and Applications to Inexact Linear Programming." *Operations Research*, 21, 1154-1157.

[19] van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models." *Annals of Statistics*, 42, 1166-1202.

[20] Wainwright, M. J. (2015). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. University of California, Berkeley.

[21] Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.

[22] Wooldridge, J. M. and Y. Zhu (2017). "Inference in Approximately Sparse Correlated Random Effects Probit Models." Forthcoming in *Journal of Business and Economic Statistics*.

[23] Zhang C.-H. and S. S. Zhang (2014). "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 217-242.

[24] Zhang, X. and Cheng, G. (2017). "Simultaneous Inference for High-Dimensional Linear Models." *Journal of the American Statistical Association - Theory & Methods*, 112, 757-768.

[25] Zhu, Y. and J. Bradic (2017). "Breaking the Curse of Dimensionality in Regression." arXiv:1708.00430.

[26] Zhu, Y. and J. Bradic (2018). "Linear Hypothesis Testing in Dense High-Dimensional Linear Models." *Journal of the American Statistical Association - Theory & Methods*, 113, 1583-1600.